QA
76
M94
SI-Ref.

M U D P I E no. 16

Museum and University Data, Program and Information Exchange

MUSEUM COMPUTER NETWORK

Under the date of February 8, 1971, David Vance(Vice Chairman of Administrative Committee, Museum Computer Network, 27 West 53rd St, New York, NY 10019) distributed a summary and history of the Network plus a series of replacement sheets to be inserted in the document entitled "Structure and Content of a Museum Data Bank" (see MUDPIE no. 11). The review of past and present condition of the Network is worth reading, and we would reproduce it here if it were not quite so long. I recommend it to MUDPIE readers. Vance has distributed additional copies of the structure and content paper to those who did not attend the meeting in New York last year, so I would guess he has copies available for others who would like to have it.--JAP.

COMPUTER CYTOLOGICAL DIAGNOSIS

BioSciences Information Service has recently announced the availability of "standard" profiles, which are designed to provide all the benefits of an individualized or customized profile of current literature to a group of scientists rather than just to one. Procedures used in preparing the profile are identical whether it is customized or standard, but the latter is more economical because the users are sharing the costs. The profile is used to search the complete index files of each issue of Biological Abstracts and the Bioresearch Index, producing a printout of the essential bibliographic information for selected items, which is mailed to all subscribers once a month. The only standard profile currently available of interest to MUDPIE is one on Computer Cytological Diagnosis. Fifteen others have been announced, however, and more will be added in response to indications of need. The annual subscription for a standard profile is $50.00 (one-half the cost of a customized service). If interested, contact R and D Department, BioSciences Information Service, 2100 Arch St, Philadelphia, PA 19103.--JAP.

SMITHSONIAN
MAY 7 1971
LIBRARIES

COMPUTERIZED BIBLIOGRAPHIES

The computer has been regarded by bibliographers as the Ark that will keep us all from being lost in a flood that makes the one Noah went through look like the squeezings from a wet wash cloth. The number of methods devised for storage and retrieval of literature on larger or smaller subjects is legion, and some of them work. Others, by virtue of their sheer bulk, or the diffi-

culties of coping with their complex computer output, do not serve the hoped-for purpose. How many others are there like myself who have practically abandoned use of Biological Abstracts in recent years? I used to be "faithful reader" personified for BA, but not any more--it is just too heavy (actually, here at the Smithsonian, the most recent numbers are always out to the bindery, anyway).

All of which brings us to a pair of new bibliographies just received or announced. The first is the "Dean Bibliography of Fishes 1968", prepared by James Atz, and published by the American Museum of Natural History, 1971, 512 pp. This is a "computerized bibliographic analysis of the greater part of the significant ichthyological literature published during 1968." 3501 titles are included, and the bibliography is in five parts: systematic, subject, geographic, author, and the original citations organized by serial number. Each of the first four parts refer to the fifth for full citation of a paper. The material appears in the format of a computer print-out, but it has been set in type. How anyone ever found enough time to search out the 3500 citations listed, cross-index them to the depth seen here, get the material set up in type and then proof-read it, all in two years, surpasseth understanding. But some things become clear from this obviously useful bibliography: 1) Only professional ichthyologists could have done the job at this depth and discrimination. 2) Any professional doing this cannot have had much time left for any other professional activities. 3) All other ichthyologists will benefit enormously from the effort. 4) The prices given indicate that ichthyologists will be able to afford it (see below). For us non-ichthyological types the question becomes: who among our professional colleagues will be available to devote the amount of time necessary to do the same job for us? Atz and his colleagues have demonstrated successfully (to me, at least) that a useful, thorough, indepth bibliography can be done at reasonable cost. The rest of us now have to find people who will do what Atz has done.

This is not to say that I agree entirely with everything as it has been done in this bibliography. The systematic index lists under individual taxa all papers by subject, and that probably is a useful idea. But both the subject and geographic indexes list papers by taxon, and that is not such a good idea. Anyone entering the file with an interest in a particular taxon will use the systematic index, and find all pertinent material. Anyone entering the file with an interest in a particular subject will wish to look at all papers on that subject, and perhaps 99% of the time would not need a column of scientific names mentioned in the paper to tell him that he should get it out and look at it. An enormous amount of space could be saved by deleting taxon references in the subject and geographic files, or perhaps retaining such references only above the generic level.

The entire bibliography is organized for computerized search

of the file, but this service will not be available for about five years, at which time the file should be big enough to make such search worth while. They anticipate being able to search the file for information concerning any taxon of fishes, any subject, or any combination of them.

The costs of the preparation of this volume have been given in detail, and they are so significant to others thinking of such operations that I quote them here in full:

Computer costs.....	$ 11000.
Printing and composition.....	16000.
Distribution of 3000 copies.....	1200.
Supporting staff (3 people).....	24000.
Office supplies.....	1000.
Total	53200.

To continue their operation, they will need 3000 subscriptions at $18.00 each, or 2000 at $26.00 each. They are offering the next volume (1969) at $10.00, which will cover only production costs.

The second bibliography is called "Computer Produced Regional Bibliography (Eastern Mediterranean, Suez Canal and Red Sea)," and has been prepared by Ruth Paldi, Sea Fisheries Research Station, Haifa, Israel. Bulletin no. 57, of the Station, describes the bibliography in considerable detail, and would presumably be available upon request from Dr. Paldi. There will be three series in the bibliography, with "A" covering 1945-70; "B" covering 1900-45; and "C" covering before 1900. Each series will include about 10 volumes, and the price per series will be $400. From the announcement I have read, it would appear that, although the entire bibliography is computerized, there is no contemplated procedure to make the stored files available for search and retrieval.-JAP.

INFORMATION SYSTEMS SURVEY
"GIS"
(Generalized Information System)

GIS is a large generalized information processing system which operates in either batch or teleprocessing mode. It provides a high level user language for query and file maintenance and in addition to Boolean selection logic, permits specification of complex arithmetic and logical processing. The system can operate on sequential and indexed sequential files and supports a limited hierarchical data organization. Query may be performed on multiple files simultaneously.

The system is an IBM program product which operates on the IBM 360 under full OS. It rents for $1500 per month.

FILE ORGANIZATION -- GIS operates on variable length logical records composed of fixed length segments of information. Up to 255 different segment types may be defined for a file and these are organized hierarchically with multiple segment types permitted at the lowest level. Up to fifteen levels of hierarchy are permitted. Each segment type is composed of a predefined collection of fixed length fields.

Files may be stored and searched sequentially on tape or disk or indexed sequentially using disk. Records stored using the sequential access method may be variable length. Those stored using indexed sequential access are stored physically as fixed length records by padding out each record to a maximum size.

Each file is described to the system through a Data Definition Table (DDT) which defines the physical characteristics of each field and the logical arrangement of fields into segments and segments into records. It also defines error conditions for the content of each field so that automatic error analysis may be performed by GIS in the course of file generation and maintenance.

FILE CREATION -- GIS files are created by the system from fixed fielded input on cards, tape or disk based upon Data Definition Tables previously stored in the system. It is necessary to provide a DDT for both the master file being created and the transaction file being used as input. Complex error detection procedures may be invoked automatically based upon specifications in the DDT.

FILE MAINTENANCE -- A GIS file may be maintained by adding or deleting complete records, adding or deleting segments from an existing record or changing individual field values. Error detection procedures are similar to those available during file creation.

Two types of file maintenance are available. The first is largely data directed and requires only a minimum of language specification to control it. Here, operations are straightforward and are based upon the content of each transaction record.

The second is language directed and may involve complex logical decisions and arithmetic calculation to determine the operation to be performed or to derive a data value. This mode may also utilize a transaction file. It may only be used to change field values.

During file maintenance multiple transaction files and multiple master files may be involved in a single run.

RETRIEVAL -- The user communicates with GIS through a high level English-like language in which he defines his selection, processing and output requirements. Depending on the complexity of the operations to be performed, the language may be very simple or quite complex and sophisticated. In all cases, the language is data independent in that the physical characteristics of the data are deduced by the

system from the DDTs and the user may refer to data by name only.

Search is performed either sequentially for sequential files or, for indexed sequential files, the search may be sequential through the entire file or by index if the file key is a necessary selection criterion. Search order is under the control of the user.

A query search may involve multiple files with data from a record selected in one file being used as the selection criteria for another file.

Data records selected during a search may be sorted for output on any fields as key.

Output capability ranges from simple list output, which is a tabular arrangement of fields formatted by the system, to complex reports whose exact specifications are provided by the user. The simple list is invoked with minimum language. Specification of a complex report format requires language analogous to a report program generator.

Multiple reports can be produced from a single procedure by creating hold files and processing each in a separate sub-procedure to produce the report.--Harriet R. Meadow.

"TAXIR" REVISITED

(The following statement concerning TAXIR was prepared by two of the original designers of the system to correct what they felt might be erroneous impressions which could be derived from the discussion of the system in our continuing "Information Systems Survey," by Harriet Meadow -- MUDPIE # 14, p. 3.--Ye ed.)

TAXIR is a powerful, completely generalized information storage and retrieval system, based on a novel retrieval technique which involves direct Boolean calculations on the stored data rather than the traditional linear file search. This technique results in significant savings in both machine storage space and machine execution time. The user addresses the system in a high level language (resembling English), which permits him complete control over building, updating, and querying data banks.

The input data may be presented to the system in either fixed-field or free-field format, and these formats may be changed by the user at any time. Data records may be added to or deleted from the data bank at desired intervals throughout the operative life of the data bank. A powerful correction statement provides complete file maintenance capability.

Although the first pilot version of TAXIR operated only on a small data bank that was core-resident, the system has since been extended to handle data banks resident in auxiliary core storage. There is no intrinsic obstacle to further extend-

ing this capability to data banks resident on tapes or disks, which will permit the system to handle indefinitely large data banks.

In defining data to the system, the user need not be limited by conforming to any predefined set of values or any predefined hierarchical structure. On requesting responses to queries, however, the user may instruct the system to organize the output in any hierarchical structure he desires. The system checks the data as it enters the machine for a wide variety of error conditions and prints messages to the user warning him of incorrect procedures or errors.

The query language of TAXIR is the language of Boolean algebra, fully generalized, so that the user may select subsets of the data bank for printout, using any possible Boolean combination of selection criteria. Thus, his queries may be very simple ones based on a single criterion, or may be as complex as he wishes, involving combinations of many such criteria. The use of parentheses and the operators NOT, AND, OR in Boolean expressions is permitted. Boolean expressions, therefore, may be written by the user to select a single item from the data bank, any possible combination of items from the data bank, or, indeed, the entire data bank. This latter capability permits the printing of hierarchically structured representations of the entire data bank, which in the case of museum applications can be the catalogue. Since the hierarchical organization is under the user's control, various hierarchical permutations of the catalogue may be created. For example, a catalogue ordered alphabetically on taxonomic name, a catalogue ordered alphabetically on geographical location, or a catalogue ordered numerically on catalogue number all are within the capability of the system. This same Boolean power permits the user to employ the querying language as a simultaneous key for the identification of unknowns.--David J. Rogers and R. C. Brill, Taximetrics Lab., Univ. of Colorado, Boulder, CO 80302.

RECENT LITERATURE

Awad, E. M. & Data Processing Management Association. 1970. Automatic Data Processing-Principles and Procedures. Prentice-Hall, New Jersey, 1970: xiv + 495.

Brandt, S. Statistical and Computational Methods in Data Analysis. Elsevier, New York, 1970: xii + 322.

Fraser, A. & D. Burnell. Computer Models in Genetics. McGraw-Hill, New York, 1970: 1-206.

Legendre, P. Circumscribing the concept of the genus. Taxon, vol. 20, 1971: 137-139.

Lockhart, W. R. & J. Liston, Eds. Methods for Numerical Taxonomy. American Society for Microbiology, Subcommittee on Numerical Taxonomy, 1970: x + 62 (A brief, readable, and adequate summary of techniques. Recommended--JAP.)

McDaniel, H. (Ed.). Applications of Decision Tables--A Reader. Brandon/Systems Press, Princeton, 1970: viii + 226.

McDaniel, H. Decision Table Software--A HANDBOOK. Brandon/Systems Press, Princeton, 1970: xi + 84. (A listing of available processors that will permit the use of decision tables with most varieties of computers currently on the market.)

Reyment, R. A., H.-A. Ramden & W. J. Wahlstedt. Fortran IV program for the generalized statistical distance and analysis of covariance matrices for the CDC 3600 computer. Computer Contribution 39, State Geological Survey, University of Kansas, 1969: 1-42.

Rohlf, F. J. Perspectives on the application of multivariate statistics to taxonomy. Taxon, vol. 20, 1971: 85-90.

Watson, L. Basic taxonomic data: the need for organisation over presentation and accumulation. Taxon, vol. 20, 1971: 131-136.

May, 1971
Division of Reptiles and Amphibians
National Museum of Natural History
Washington, D. C. 20560